OPEN ACCESS

University of BRISTOL

Damen, D., Leelasawassuk, T., & Mayol-Cuevas, W. W. (2016). You-Do, I-Learn: Egocentric Unsupervised Discovery of Objects and their Modes of Interaction Towards Video-Based Guidance. *Computer Vision and Image Understanding*, *149*, 98-112. https://doi.org/10.1016/j.cviu.2016.02.016

Peer reviewed version

License (if available):
CC BY-NC-ND

Link to published version (if available):
10.1016/j.cviu.2016.02.016

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Elsevier at http://www.sciencedirect.com/science/article/pii/S1077314216000709. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research
General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

You-Do, I-Learn: Egocentric Unsupervised Discovery of Objects and their Modes of Interaction Towards Video-Based Guidance

Dima Damen, Teesid Leelasawassuk, Walterio Mayol-Cuevas

Computer Science Department
University of Bristol, Bristol, UK

Abstract

This paper presents an unsupervised approach towards automatically extracting video-based guidance on object usage, from egocentric video and wearable gaze tracking, collected from multiple users while performing tasks. The approach i) discovers task relevant objects, ii) builds a model for each, iii) distinguishes different ways in which each discovered object has been used and vi) discovers the dependencies between object interactions. The work investigates using appearance, position, motion and attention, and presents results using each and a combination of relevant features. Moreover, an online scalable approach is presented and is compared to offline results. The paper proposes a method for selecting a suitable video guide to be displayed to a novice user indicating how to use an object, purely triggered by the user's gaze. The potential assistive mode can also recommend an object to be used next based on the learnt sequence of object interactions. The approach was tested on a variety of daily tasks such as initialising a printer, preparing a coffee and setting up a gym machine.

Keywords: Video Guidance, Real-time Computer Vision, Assistive Computing, Object Discovery, Object Usage

1. Introduction

Increasingly, commercial interest in wearable devices, including cameras and head-mounted displays in miniature and in fully wearable form (e.g. Google's Glass, Microsoft's HoloLens, Sony's SmartEyeglass) invited research into cognitive systems that take advantage of these platforms. Footage from wearable cameras has fuelled Internet-based video sharing sites. Interestingly, among the

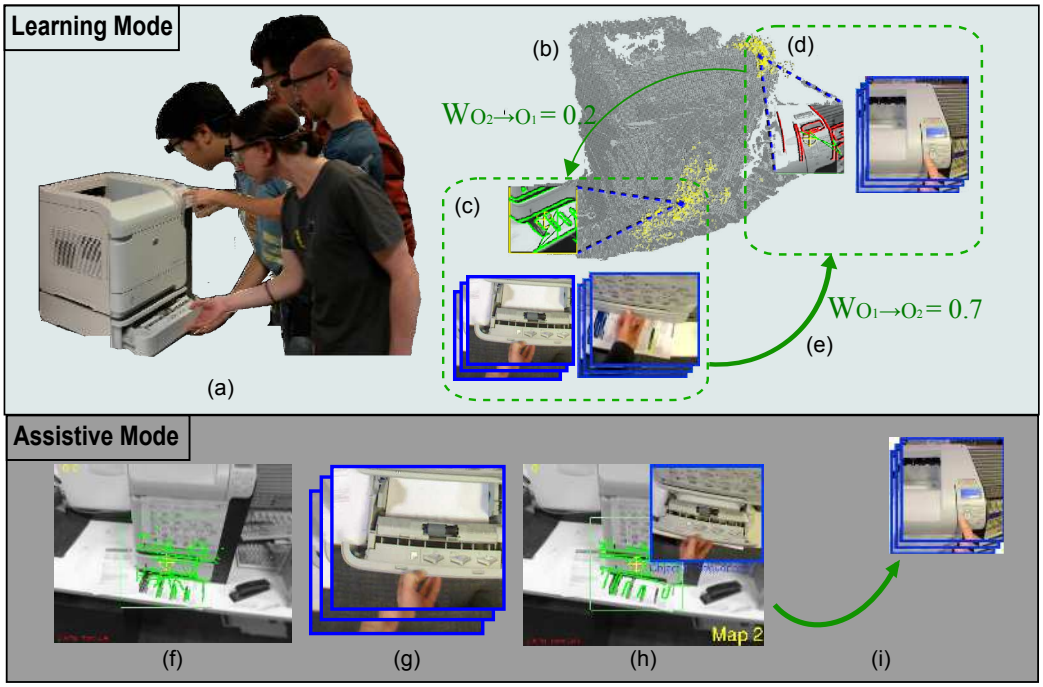


Figure 1: Given egocentric videos from multiple users (a), a map of the environment (b) and feature clustering are used to discover distinct task-relevant objects (TROs); e.g. paper drawer and keypad for the task of operating a printer (c,d). For each discovered TRO, a model is built to incorporate possible locations, appearance and usage, along with a probabilistic graph of object interactions (e). In a potential *assistive mode*, when a TRO is recognised triggered by gaze (f), a usage snippet can be chosen (g) and can be displayed to the user to provide guidance on how to use the object (h) along with the most likely object to be used next (i).

most sought after videos are *how to do* guides, accessed by people wishing to carry out tasks, from cooking to assembling furniture.

Assistance in task performance (e.g. assembly, repair) using augmented reality or video guidance has been promised for a while. One of the key limitations to realise such systems is the time consuming and evidently limiting task of authoring the content by e.g. manually segmenting and annotating videos or creating three-dimensional models that represent meaningful guidance (e.g. [45],[2]). Approaches that can discover object interactions from video input and provide guidance without the need for manual intervention would enable a wider adoption of assistive wearable systems.

Figure 1 shows an overview of the proposed *You-Do, I-Learn* approach, both the learning and the assistive modes. This work attempts, to **fully unsupervised**,

discover objects and their usage from multiple users in a common environment (Fig. 1a), then proposes a complete automatic solution for object-based guidance for discovered objects. *Note that the suggested assistive mode can potentially be implemented on a wearable device, but this is out of the scope of this work.*

In proposing the *You-Do, I-Learn* approach, we particularly focus on an egocentric view of the world, taking advantage of wearable technology, as it offers a unique perspective on object-level interactions. As opposed to discovering all objects in the environment, we focus on discovering task relevant objects. A **Task Relevant Object (TRO)** is an object, or part of an object, with which a person interacts during task performance. For example, a person operating a printer may interact with the paper drawer (Fig. 1c) and/or the keypad (Fig. 1d) while operating it. A system that aims to discover TROs would attempt to discover these objects/parts as opposed to the full machine or all of its parts. For each discovered object, we build a location model, an appearance model and collect usage snippets on how different users interacted with the same discovered object. A **usage snippet** is an automatically extracted video sequence, to reflect how an object has been used. Several usage snippets can be extracted for the same TRO as the object is used multiple times by the same or different users.

The various models of TROs can be used to provide assistive guidance. The location model guides the user to where an object can be found. The appearance model is used to recognise the object when visible in the wearable device's field of view. The collected usage snippets can be used for video-based guidance. To achieve video-based guidance on object usage, we also introduce the term **Modes of Interaction (MOI)** to refer to the different ways in which TROs are used. Say, a cup can be lifted, washed, or poured into. All these are different MOIs associated to the cup. When harvesting usage snippets for the same object from multiple operators, common MOIs can be discovered. In introducing MOIs, we distinguish between object-based guidance and task-based guidance. This is because the same object can be used in many tasks, while the ways in which one object can be interacted with are usually limited to a finite set of possible interactions. Triggered simply by gaze (Fig. 1f), the user is advised on how a TRO object can be used based on the object's current state (Fig. 1h), as well as advise the user on the most-likely object to be used next (Fig. 1i).

Section 2 presents an overview of previous attempts towards object discovery, in general and for egocentric video in particular, as well as attempts towards video-based guidance. The learning and assistive modes are presented in Sections 3 and 4 respectively. A varied dataset from coffee preparation to operating a gym machine is presented, alongside results in Section 5. Section 6 discusses building

approximate three-dimensional models for discovered objects, as a by-product of the approach. These can be used for virtual reality guidance, but this is left for future work. The paper concludes with future directions.

2. Video-Based Object Discovery and Guidance - a Review

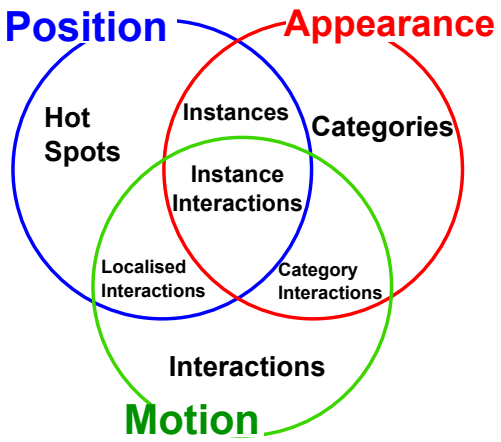


Figure 2: Using appearance, position, motion and combinations for object discovery

Object discovery refers to grouping feature descriptors into meaningful clusters that correspond to entities worth discovering. We attempt to differentiate the various ways in which entities can be discovered from video input; appearance, position and motion. Figure 2 envisages what can be discovered if each, or a combination, of these information is used in the grouping. The **position**, relative to an environment, can be grouped into hot-spots. A *hot spot* is a position at which object interaction takes place, and could correspond to objects that remain fixed relative to the environment. **Appearance** similarity is a strong cue to discover visual categories, i.e. objects that share similar appearance. When combining appearance with position, instances of objects can be discovered. In video, **motion** presents a third cue that could be employed solely to discover ego actions and interactions. When combined with instances, different manners of interaction with objects can be discovered.

The number of works attempting to use egocentric vision for tasks ranging from attention estimation to activity recognition has increased exponentially in the last decade. The reader can refer to early works [39, 40] or more recent surveys [1, 43] for a review of datasets, methods and attempted problems. In this

review, we focus on *unsupervised* approaches to object discovery (Section 2.1), task-relevant object discovery from egocentric video (Section 2.2) as well as approaches that aim to link discovery to guidance (Section 2.3). The works presented here differ from the frequent attempts to recognise objects in egocentric video from supervised training (e.g. [15, 46, 48]) which are interesting in their own right.

2.1. Unsupervised Object Discovery from Static Images

Appearance similarity, along with the objects' positions, has been deployed in various works to discover objects from a set of images or 3D scenes in an unsupervised manner.

Appearance: Appearance, of object and context, is often used to discover *categories* (e.g. [25, 49, 57]) or *instances* (e.g. [24, 21, 22, 51]) of an object based on visual similarity. When attempting to discover categories, most works assume a collection of images where *common* features correspond to one category, with or without spatial consistency (e.g. [25, 49]). These approaches are only semi-supervised as collecting images belonging to a category is needed. In [57], Tuytelaars *et al* review and compare recent works in category-based discovery from a dataset of images.

Unsupervised instance discovery, similar to our work, has also been attempted from a set of images and 3D scenes [24, 21, 22]. In [24], 3D object segments are discovered from a dataset of indoor RGB-D scenes. Several measures assist segmentation: compactness, symmetry, local convexity, global convexity, smoothness and recurrence in multiple scenes. Object segment hypotheses are accordingly ranked to come up with a final set of discovered objects. In [21], a data-driven objectness measure is proposed where a segment is compared to a database of general object segments. In [22], colour, texture and shape-based features are used to construct a network of finely-segmented regions. Segments are then iteratively grouped and refined until the algorithm converges to discovered objects. While a very interesting approach with promising results, [22] assumes that objects of daily living are moveable. A computer screen, for example, needs to be moved to a different background to enable its discovery. Many objects of daily living such as a coffee machine or an electric socket remain fixed to their surroundings. The approaches in [21, 22] also assume the dataset contains at least one instance of an object of interest per image. When using video as input, a significant number of frames might not contain TROs as the user roams around an environment. Moreover, these approaches are processed offline after the dataset is collected.

Position: Position information has been used solely in [18] to discover objects, by aligning two point clouds and identifying changes in location that correspond to objects that have been placed or removed. In [37], the disappearance of features in a position is used as a cue to discover objects from RGB-D images collected using a mobile platform. This assumption is also considered by an earlier work [51] which uses multiple cameras to build a depth map.

Appearance and Position: Combining position and appearance cues has been used to enable discover objects [52, 5]. In [52], sensor tracking enables constructing a model of objects placed on a planar surface. Segments are combined from multiple scenes using appearance matching of interest points.A database of models is used to refine the reconstruction of discovered objects. In [5], RGB-D images collected from a robot in a common environment are first separated into discrete locations (rooms, in their case), then appearance and depth data are clustered to extract instances. The approach assumes that all objects are placed on a planar surface (e.g. table-top) and employs a prior on the object's shape and size.

2.2. Unsupervised Discovery of Task-Relevant Objects from Egocentric Video

Egocentric video introduces two new sources of information to object discovery; motion and attention. Motion is the result of the wearer's self-motion or objects in an environment. Egocentric video shows a vantage viewpoint to objects the person attends to. This section reviews works in egocentric video analysis towards, or related to, unsupervised discovery of objects.

Motion: In egocentric video, motion descriptors have been proposed for action recognition, either full-body actions [26] or object interactions [54, 55]. In [58], unsupervised discovery of object interactions is attempted, by clustering unlabelled video snippets representing actions. The problem is formulated as a linear program and a solver is used to find the optimal clustering, using the earth mover's distance measure. Though unsupervised in nature, the approach assumes the number of tasks (i.e. the number of clusters) is known, based on the knowledge that the people in the dataset perform a pre-specified set of tasks. The approach compares K-means, Kernel K-means as well as convex and semi-nonnegative matrix factorisation.

Appearance, Motion and Attention: As opposed to discovering all objects, several works focus on discovering objects with which the person interacts, whether towards object discovery or action recognition.

In a recent work, Bolanos proposed a semi-supervised approach to discover objects from wearable sensors [3]. The video is uniformly sampled into a sparse set of images at $\frac{1}{60}$fps. Given partial labelling, objectness measures along with

Convolutional Neural Networks (CNN) are used as features with iterative clustering. Clustering is evaluated, using silhouette coefficients, to decide on the discovered objects.

In [13], an interaction is identified by the change in appearance of the object before and after the action is performed. Foreground segmentation is used, followed by extracting the hand-held object regions. Unsupervised clustering of object segments enables modelling the change in the object's appearance.

In [32], objects of 'importance' are segmented from egocentric video sequences using appearance and motion features. Segmentation is based on the similarity to 'segments of importance' from a manually labelled training set, collected via crowd sourcing.

Accordingly, common approaches to discovering TROs in egocentric vision include i) segmenting the area connected the user's hand [14, 13, 32], ii) extracting foreground regions through frame stabilisation or scene planarity assumptions [47, 55] or iii) detecting 'object-like' regions [36, 3]. The first two approaches are only able to segment objects while being manipulated, during which objects could be heavily occluded by the hand. In the second approach, fixed objects like a sink tap or a coffee machine, which can be quite crucial to a task, are ignored. In the third approach, 'object-like' regions can focus on salient rather than used objects.

Very few systems exploit the high quality and predictive nature of eye gaze fixation. Its anticipatory nature allows estimating which object will be used next [30, 29]. Gaze has been successfully used to assist action and activity recognition [12, 35, 44, 38] and supervised object recognition [53, 11].

2.3. Unsupervised Video-Based Guidance

Unsupervised extraction of video snippets from a continuous egocentric video has mostly targeted video summarisation [36, 32]. The earliest example we could trace of segmenting video snippets for guidance is the work of Kang and Ikeuchi in 1994 [23] that uses stereo visual data and other sensors, and focuses on tracking hand motion. The extracted snippets are used for guidance of robotic arms during grasping. Similarly, the work of Mayol and Murray in 2005 [41] automatically detects keyframes of interactions with objects from a shoulder-mounted camera, towards event-based summarisation.

For human user assistance, Hashimoto et al. [17] proposed view sharing of video from wearable cameras to guide novice users. Their work does not require unsupervised segmenatation of video guides but focuses on live sharing of egocentric views. In [16], instructional videos are projected onto an AR display for task

guidance. While manually edited instruction video clips are employed, the system paces the instruction clips to match the status of the performed task. In [45], automatic extraction of snippets is performed using novelty detection. Video clips are extracted based on the distance between consecutive frames. The work also discusses overlaying the segmentation videos onto the scene in real time.

Up to our knowledge, this manuscript presents the first attempt to close the gap between object discovery and video-based guidance in a fully unsupervised way. The manuscript builds on our previous works towards offline [9] and online [8] discovery of task relevant objects, with additional novel contributions:

- A generalised formulation for the problem of discovering task-relevant objects from a sequence of egocentric images.
- An improved online discovery algorithm compared to the one proposed in [8] with superior performance. The new algorithm uses a Gaussian Mixture Model (GMM) to represent each discovered object as opposed to a threshold over Euclidean distance from [8].
- Previously unpublished comparison of the online approach to offline discovery of TROs.
- Building a graph of object interactions, which can provide guidance on which object to use next in a sequence of object interactions.
- A detailed explanation of how subject annotations can be used to generate ground-truth of task-relevant objects and their usages.

Moreover, the paper provides further details on both the offline and the online algorithms. The approach is explained next; first the learning mode (Sec. 3) then the assistive mode (Sec. 4).

3. You-Do, I-Learn: Learning Mode

During learning, Task-Relevant Objects (TRO) need to be discovered (Sec. 3.1) and a model to be built for each object (Sec. 3.2). For each discovered TRO, usage snippets are automatically collected showing multiple people interacting with the same object. These usage snippets can be analysed to discover the various Modes of Interaction (MOI) in an unsupervised manner (Sec. 3.3). Sequences of object interactions can also be discovered, highlighting strong temporal dependencies (Sec. 3.4).

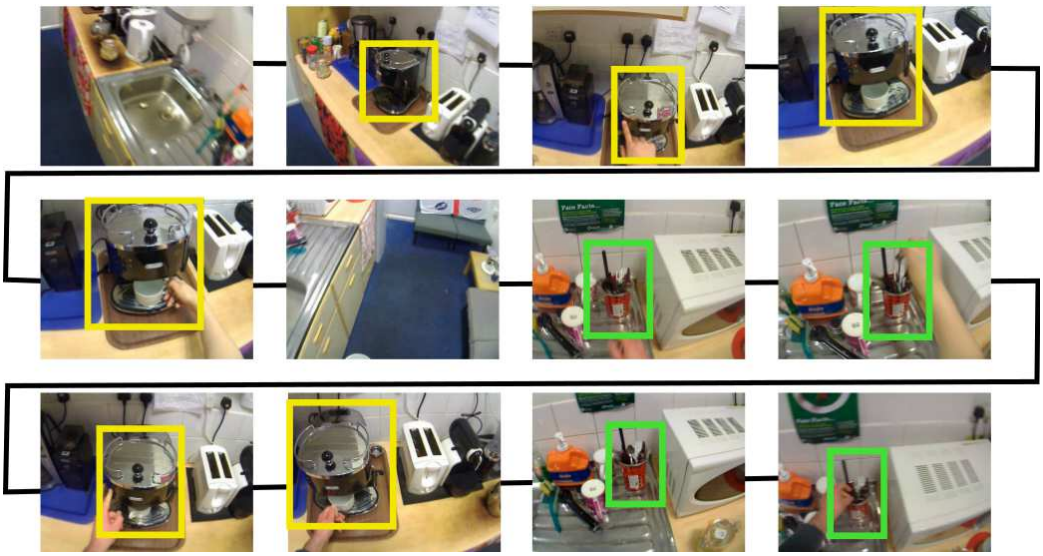


Figure 3: Given a sequence of images from egocentric views, the objective is to group image parts into TROs, based on the assumption that one image part at most is task-relevant in each image. In this example, two TROs are shown.

3.1. Discovering Task Relevant Objects (TRO)

We first present a formulation for the problem of discovering task-relevant objects from egocentric video. Given a sequence of egocentric images $\{I_1, .., I_T\}$ collected from multiple operators around a common environment, TRO discovery is the process of finding K TROs, $\{O_k; 1 \leq k \leq K\}$, where the number of objects K is not known *a priori*. Assume $\Omega(I_t)$ is a part of the image I_t (e.g. a segmentation or a bounding box within I_t), each discovered TRO O_k is a set of image parts from the sequence.

$$O_k = \{\Omega(I_t); 1 \leq t \leq T\} \tag{1}$$

In this formulation, we make the assumption that at most one task-relevant image part is present within each image I_t. The notion of 'at most' handles cases when the person is not actively interacting with any object in the environment. During interactions, only one image part is task-relevant. The person could be interacting with multiple objects, for example placing one object on top of another, yet the attention is believed to shift between these objects [20]. This assumption simplifies the discovery of TROs without much loss in generality. Accordingly, the sets of image parts representing discovered TROs $\{O_k\}$ are believed to be

disjoint and form a subset of all images in the sequence. Figure 3 shows a visual representation of TRO discovery formulation.

Given the above formulation, we next propose two approaches to TRO discovery, one is offline assuming all sequences from multiple users are collected prior to the discovery. The second approach is online, and thus is scalable to multiple users and different TROs. For image parts $\Omega(I_t)$, we only report results on using bounding boxes. The proposed online and offline approaches are applicable to segmentations, but this is left for future work. We compare two techniques to suggest a bounding box in an image that can contain a TRO. The first $\Omega_c(I_t)$ crops the image around the centre. Given a glass-mounted camera, it is expected to have the object of interest at the centre of the frame during interactions. We compare this approach to gaze fixation $\Omega_g(I_t)$ where the image is cropped around a known gaze fixation. Using a wearable gaze tracker, we filter saccades using the velocity-based approach from [50], where the average angular velocity over a sliding temporal window is considered a saccade if it is greater than $100^\circ/sec$, and is thus discarded. Note that we do not use image-based saliency to find the image part $\Omega(I_t)$. There are potentially many visually salient objects in an environment that are not interacted with. Our interest is to discover only those objects, whether visually salient or otherwise, with which one interacts.

To describe image parts, we use position and appearance features as well as their combination following Fig. 2:

- **Position:** The Image I_t is positioned relative to the scene using sparse Simultaneous Localisation and Mapping (SLAM) [27]. A triangular tessellation of tracked interest points is built, similar to [56]. Given the 6D pose of the scene camera, a 3D ray connects the centre of the image part $\Omega(I_t)$ to a point in the scene. Using the tessellation, the 3D position of the intersection point is calculated using linear interpolation. *Using position features solely enables discovering static TROs. Moveable objects, observed in different locations will be discovered as separate TROs.*

- **Appearance:** To represent appearance, Histogram of Oriented Gradients (HOG) [6] is calculated over sub-windows within the image patch $\Omega(I_t)$. In offline TRO discovery, Bag of Words (BoW) representation is used for appearance information. In online TRO discovery, HOG features are used as appearance features directly. This is because BoWs require either a representative prior training sequence or an adaptive approach that can merge and introduce new words. Generalisation of an adaptive BoW approach to the variety of locations and tasks we report in the experiments section would not

be trivial. *Using appearance features solely enables discovering moveable objects, as clusters combine observations of a similar appearance. Static objects though could be separated into multiple clusters if varying viewpoints are observed.*

- **Combining Position and Appearance:** When combining position and appearance features, the normalised affinity matrices are summed with equal weighting in offline TRO discovery. The features are simply combined for online TRO discovery. *By combining position and appearance features, static objects can be discovered and moveable objects can be combined using appearance feature similarity.*

We also compare to results that accumulate features over a sliding window w centred around each image $(\Omega(I_{t-\frac{w-1}{2}}), .., \Omega(I_t), .., \Omega(I_{t+\frac{w-1}{2}}))$. In the experiments section, we test features that use position, appearance and their combination, over a sliding window and the two image part methods. We use the term f_t next to refer to the feature vector representing an image part where,

$$f_t = (F(\Omega(I_{t-\frac{w-1}{2}})), .., F(\Omega(I_t)), .., F(\Omega(I_{t+\frac{w-1}{2}})) \tag{2}$$

and $F(\Omega(I_t))$ is the feature descriptor for the image part $\Omega(I_t)$.

3.1.1. Offline TRO Discovery

Offline TRO discovery refers to the attempt to discover all TROs after the dataset is fully collected. The sequencing of images is thus discarded and a data point $x_i = f_t$ refers to the descriptor of an image part in the dataset. We compare k-means clustering to spectral clustering from Ng *et al.* [42]. These approaches were compared in [57] for a known number of object categories.

Unsupervised discovery, like other grouping problems, suffers from the dilemma of model selection (i.e. the optimal number of groups). Most previous approaches assume the number of groupings is known *a priori* [25, 57] to avoid the complexity. We propose estimating the optimal number of clusters $\hat{K}$ using the Davies-Bouldin (DB) index [10]. For an object O_k with n_k data points $\{x_i; i = 1..n_k\}$ assigned to this cluster, and μ_k is the mean of these data points, the intra-cluster distance S_k can be measured as (Euclidean distance used):

$$S_k = \sqrt{\frac{1}{n_k} \sum_{i=1}^{n_k} ||x_i - \mu_k||_2} \tag{3}$$

The inter-cluster distances between two objects O_k and O_j is measured as $M_{kj} = ||\mu_k - \mu_j||_2$. The cluster similarity measure $R_{kj} = \frac{S_k + S_j}{M_{kj}}$ is used to calculate DB index,

$$V_{DB}(K) = \frac{1}{K} \sum_{k=1}^{K} \max_{j \neq k} R_{kj} \tag{4}$$

The optimal number of clusters is calculated to be

$$\hat{K} = \arg\max_{K} V_{DB}(K) \tag{5}$$

Recall that some images do not contain a TRO (Fig. 3), while clustering assigns a label for each data point. We assign a probability to each cluster being a TRO as the ratio of the number of points in the cluster to the total number of points,

$$p(S_k) = \frac{n_k}{\sum_{j=1}^{\hat{K}} n_j} \tag{6}$$

Clusters are refined by removing the furthest $\beta\%$ of points in the cluster from the mean μ_k. The refinement threshold, β, was set to 75 in all experimental results.

3.1.2. Online TRO Discovery

To discover objects in an online manner, image parts are clustered as they are collected and clusters are incrementally updated. An approach for online TRO discovery should iteratively cluster image parts of the same object as the object is used by multiple operators, whether in the same or a different location.

In proposing an algorithm for online TRO discovery, we rely on the assumption that *consecutive similar image parts* $(\Omega(I_{t-\xi+1}), .., \Omega(I_t))$ *indicate an observation of a task-relevant object (TRO)*. We thus define a TRO O_k as a collection of 'at least' ξ *consecutive and similar* image parts. The notion of similarity relies on the features used. For example, when f_t is the 3D position of image part $\Omega(I_t)$, then at least ξ spatially-close consecutive image parts are labelled as a TRO. Alternatively when f_t is the appearance of image part $\Omega(I_t)$, then at least ξ consecutive image parts of similar appearance enable discovering a TRO.

Two consecutive image parts, $\Omega(I_t)$ and $\Omega(I_{t-1})$ belong to the same object if $||f_t - f_{t-1}|| < \epsilon_1$, where ϵ_1 is the threshold selected to accept clustering consecutive image parts and $||.||$ is the Euclidean distance (Algo. 1 L. 9-14). The strict consecutive constraint between t and $t-1$ can be relaxed to allow proximity within a sliding window. The mean and covariance of O_k are updated incrementally as

input : Image parts and feature vectors $\{(\Omega(I_t), f_t)\}; t = 1..T$
output: TROs $\{O_k; 1 \leq k \leq K\}$ where $O_k = (\{\Omega(I_t)\}, \Phi_k)$ and $\Phi_k = \{(\theta_{ki}, \mu_{ki}, \Sigma_{ki}); i = 1..L_k\}$

K = 0
$candidate$ = 0
for $t = 1..T$ **do**
  find closest cluster k: $\min \arg_k \sum_{i=1}^{L_k} \theta_{ki} ||f_t - \mu_{ki}||_{\Sigma_{ki}}$
  if $\sum_{i=1}^{L_k} \theta_{ki} ||f_t - \mu_{ki}||_{\Sigma_{ki}} \leq \epsilon_2$ **then**
    $l = \min \arg_l ||f_t - \mu_{kl}||_{\Sigma_{kl}}; 1 \leq l \leq L_k$
    Update θ_{kl}, $\mu_{kl}(Eq\ 7), \Sigma_{kl}(Eq\ 8)$
  else
    if $||f_t - f_{t-1}|| < \epsilon_1$ **then**
      $candidate$ = $candidate$ + 1
      if $candidate \geq \xi$ **then**
        $K = K + 1$
        $L_K = 1$
        Calculate μ_K and Σ_K
    else
      $candidate$ = 0
  if $\min_{j \neq k} d_B(O_k, O_j) < \epsilon_3$ **then**
    $L_j = L_j + 1$
    $\mu_{jL_j} = \mu_k$
    $\Sigma_{jL_j} = \Sigma_k$
    Calculate mixture components θ_j
    Delete O_k (objects merged)
    K = K - 1

Algorithm 1: Proposed algorithm for *online* TRO discovery

further image parts are located within the threshold ϵ_1. Equations 7 and 8 show the incremental update for the mean and covariance of a O_k.

$$||f_t - f_{t-1}|| < \epsilon_1 \quad \rightarrow \quad \mu_t^k = \frac{\mu_{t-1}^k \times (n_t^k - 1) + f_t}{n_t^k} \tag{7}$$

$$\rightarrow \quad \Sigma_t^k = \frac{n_t^k - 2}{n_t^k - 1}\Sigma_{t-1}^k + \frac{1}{n_t^k}(f_t - \mu_{n_t^k - 1}^k)^T(f_t - \mu_{n_t^k - 1}^k) \tag{8}$$

where μ_t^k is the mean, Σ_t^k is the covariance matrix and n_t^k is the number of image parts within O_k at time t.

Attention is believed to have moved to another object when $||f_t - f_{t-1}|| \geq \epsilon_1$. At a future point in time $t + \rho$, further image parts $\Omega(I_{t+\rho})$ can belong to the same TRO O_k if it is within ϵ_2 standard deviations from the TRO k using the Mahalanobis distance (Algo. 1 L. 4-5). This clustering method does not predefine the size of the clusters. When using position as a feature, it enables both small-sized and large TROs to be discovered.

The algorithm enables combining observations of the same object, in different locations, into the same cluster (Algo. 1 L. 17). When using appearance-based similarity, observations of moveable objects are grouped together. Two clusters (μ_t^j, Σ_t^j) and (μ_t^k, Σ_t^k) are merged if the distance measure d_B is below a threshold ϵ_3. We use Bhattacharyya distance over appearance features for merging clusters. As multiple observations of a moveable object O_k are not necessarily close in spatial location, a Gaussian Mixture Model (GMM) $\{(\theta_i, \mu_i, \Sigma_i), i = 1..L_k\}$ is used to represent the location model where θ_i is the mixture component of the Gaussian i and L_k is the number of Gaussians in the GMM for object k (Algo. 1 L. 18-23). A new Gaussian is added to the GMM every time an object of similar appearance is found in a new position.

3.2. Building Models of TROs

Section 3.1 proposed an offline as well as an online approach for discovering task relevant objects from egocentric video. For each discovered object O_k, we build three models that encapsulate the object's location Φ_k, appearance A_k as well as its usage U_k. As the models are built from multiple operators with different heights and interaction behaviours, they give a good representation of the object (e.g. Fig. 4). These models can enable a broad range of potential assistance to users. The location model guides the user to where an object can be found. The appearance model is used to recognise the object when seen again. The collected usage snippets are used for video-based guidance in Section 3.3. We next detail how the various models can be built for each discovered TRO.

Location Model Φ_k: The location model represents the position and extent of the object using a Gaussian Mixture Model (GMM) Φ_k; a single Gaussian for a fixed object and a multi-variate Gaussian for moveable objects. In the offline approach, position features are clustered and the DB-index is used to decide on the number of Gaussians in the GMM. In the online approach, a new Gaussian is introduced for every observation in a new location. The likelihood of the object's position is then,

$$P(f_t|O_k) = \sum_{l=1}^{L_k} \theta_{kl} e^{[-\frac{1}{2}(f_t-\mu_{kl})^T \Sigma_{kl}^{-1}(f_t-\mu_{kl})]} \tag{9}$$

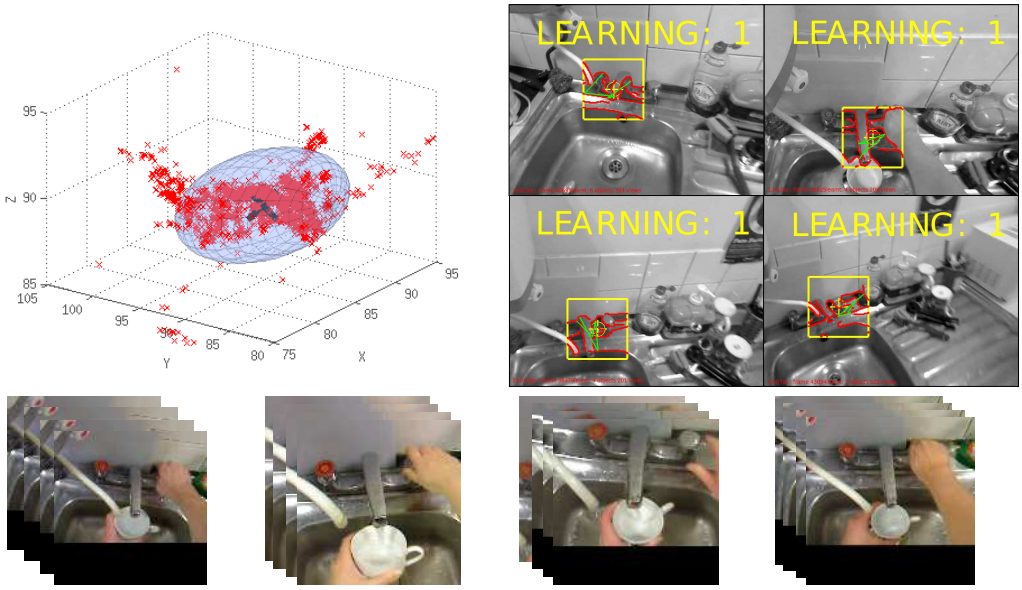


Figure 4: For a discovered TRO (tap): multiple users enable modelling the object's position Φ_k (top-left) learning varying views in the appearance model A_k (top-right) and gathering different usage snippets U_k that show interactions with the same object (bottom).

Appearance Model Φ_k: For a view-based appearance model, we use the real-time method from [7] for learning novel views of the object. This method is particularly helpful for online learning as it is scalable and works in real-time. It is shape-based and thus is particularly suitable for texture-minimal objects, many of which are present in indoor environments.

Usage Model U_k: The consecutive image parts clustered into one TRO are combined into video snippets indicating how a TRO was used by multiple operators. Given consecutive image parts $\{\Omega(I_t), \Omega(I_{t+1}), \Omega(I_{t+\rho})\}; \rho \geq \xi$ belonging to the same TRO, a **usage snippet** u_i^k is a video clip formed by combining the sequence of image parts. Notice that when using gaze fixations Ω_g, interpolation is needed when gaze information is missing. The collection of all usage snippets $U_k = \{u_i^k\}$ shows different ways in which O_k was used or interacted with. These usage snippets are further analysed in Section 3.3 for discovering the various modes of interaction with the same TRO.

3.3. Finding Modes of Interaction (MOI) for TROs

For discovered TROs, we find common MOIs for each TRO by analysing usage snippets, each representing a sample usage. The collection of all usage

snippets $U_k = \{u_i^k\}$ shows different ways in which O_k was used. Position and appearance information of all frames in u_i (superscript k removed for simplicity) are the same features used for discovering objects. These are augmented with motion information collected using the Histogram of Optical Flow (HOF) descriptors around 3D Harris points [31] to encode the interaction with the object.

We also use a *temporal pyramid* to encode the descriptors. At each level l, the snippet is split into l equally-sized temporal segments, and the descriptor is calculated for each segment. The temporal pyramid could potentially separate MOIs that differ in their temporal ordering, such as opening and closing. A one-dimensional representation of the temporal pyramid formulates the descriptor $d(u_i)$. Clustering then follows (as in 3.1.1) to find the MOIs.

Each cluster is represented by the snippet $\hat{u_j}$ closest to the centre of the cluster μ_j (i.e. mean snippet),

$$\hat{u_j} = \underset{u_l \in MOI_j}{\arg\min} ||d(u_l) - \mu_j||; \quad \mu_j = \frac{1}{|MOI_j|} \sum_{u_l \in MOI_j} d(u_l) \tag{10}$$

and the confidence in a cluster being a common mode of interaction is represented by the percentage of snippets within that cluster $p(MOI_j)$,

$$p(MOI_j) = \frac{|MOI_j|}{|U_k|}; \qquad p(MOI_j) \geq \lambda \tag{11}$$

A threshold λ can be used to select common MOIs such that $p(MOI_j) \geq \lambda$.

3.4. Graphs of Object Interactions

Following the discovery of TROs, it is also possible to model, in an unsupervised way, the sequence of object interactions towards modelling tasks or simply discovering strong links between object interactions. For example, after using the tap, the user is likely to follow that by interacting with the drainer or with a towel. These strong links between objects can be automatically discovered from sequences of multiple users. While a more complex interaction model can be targeted, we employ the first-order Markovian assumption. We model TRO interaction sequences by a graph-based representation.

For all discovered objects $\{O_k; k = 1..K\}$, a *complete* directed graph G is constructed so each TRO is represented by a node and the weight $W_{O_k \to O_j}$ of the directed edge $O_k \to O_j$ represents the probability of interacting with object O_j directly after having interacted with object O_k. Note that we loosely define interaction as attending or looking at an object. The edge weights are initialised

with a small value α. Temporal transitions from one discovered object to another are accounted for, followed by edge-weight normalisation.

4. You-Do, I-Learn: Assistive Mode

In the assistive mode, the location models $\{\Phi_k\}$, the appearance models $\{A_k\}$, the usage snippets $\{U_k\}$, the various modes of interaction $\{MOI_k\}$ as well as the graph of object interactions $G_{K\times K}$ are used to provide a recommendation of how an object can be used, as well as what object to use next.

To provide guidance, the object with which the user is attempting to interact should be recognised. In a test image I_t, the image part $\Omega(I_t)$ is compared to the discovered TROs. Upon recognition of a TRO O_k, video-based guidance can be provided by showing one of the possible MOIs, that is most relevant to the task or the object status. The *help snippet* $h_t = u_k \in U_k$ is chosen from the possibly many *usage snippets* featuring the TRO. We choose the *help snippet* h_t as a usage guide at time t such that the appearance of the first frame in the snippet, is closest to the recognised view.

$$h_t = \arg\min_{u_j} ||A^{1st}(u_j) - A(\Omega(I_t))|| \tag{12}$$

where A^{1st} is the appearance of the first frame in the snippet. If the object changes state, the initial appearance is a good indicator of which usage snippet to show. An additional advantage is to avoid showing a snippet observing the object from a different viewpoint, so the user can easily map what they see to what they could do. A *help snippet* is displayed each time a new object is detected, aiming to provide automatic assistance for novice operators.

The graph of object interactions $G_{K\times K}$ can be used to estimate the object to be next manipulated. The assistive mode would recommend the object to be used next, so that

$$\hat{j} = \arg\max_j p(O_j|O_k) = \arg\max_j W_{O_k\rightarrow O_j} \tag{13}$$

The location model for the recommended object $O_{\hat{j}}$ can be used to suggest where the object is likely to be found (Eq. 9),

$$\hat{l_j} = \arg\max_{f_t} P(f_t|O_{\hat{j}}) \tag{14}$$

Recommending a help snippet, the object to be used next as well as where that object can be found are based on correctly recognising that the user is attending a TRO. We base the assistive mode on gaze fixations $\Omega_g(I_t)$ and investigate two approaches for recognising the TRO,

1. **Using the location models** $\{\Phi_k\}$: a TRO is recognised based on Eq. 9 so that

$$k = \arg\max_k p(f_t|O_k); \qquad p(f_t|O_k) \geq \lambda \tag{15}$$

2. **Using the appearance models** $\{A_k\}$: this assistive mode does not require a map of the environment or tracking of the camera relative to the environment. Given the image part $\Omega_g(I_t)$, the appearance model is used to recognise the viewed object, from the set of appearance models. By using the combination of fixed paths and a hierarchical hash table, object recognition is scalable, and can reliably detect objects at frame rate [7]. The descriptor is affine-invariant, and the method is tolerant to a level of occlusion but is also view-dependant. Figure 5 shows the method learning (left column) and subsequently recognising (right column) objects from our experiments.

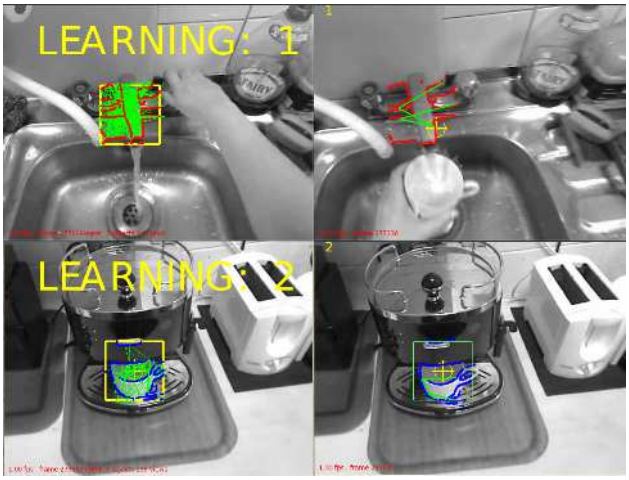


Figure 5: During discovery (left), edges within $\Omega(I_t)$ are captured as object views, and represented using affine-invariant descriptors [7]. These are used to detect objects around the gaze point in real-time (right).

5. Experiments and Results

Setup & Dataset: The wearable gaze tracker hardware (ASL Mobile Eye XG [28]) consists of two cameras sharing a half-mirror, one looking at the scene and another looking at the eye. After calibration, the scene images are synchronised with, if available, 2D gaze points. Six locations were chosen: kitchen (K),

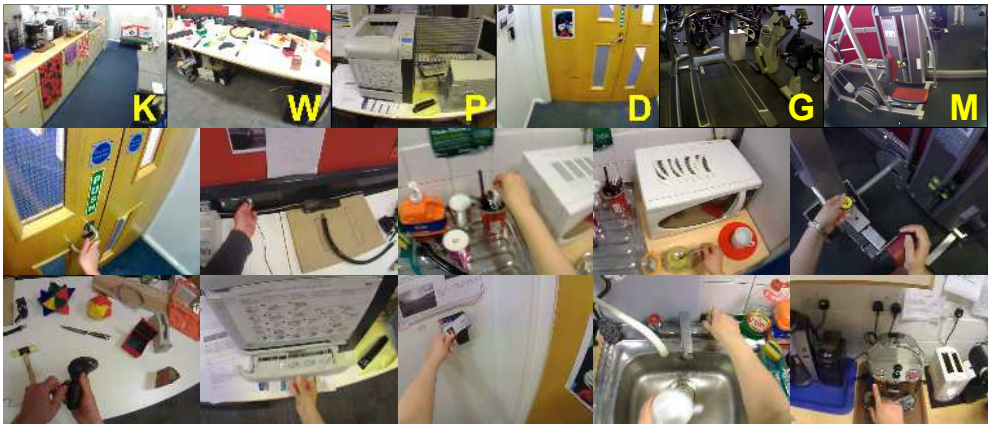


Figure 6: Sample images from the Bristol Egocentric Object Interactions Dataset (BEOID).

workspace (W), laser printer (P), corridor with a locked door (D), cardiac gym (G) and weight-lifting machine (M) (Fig. 6). For the first four locations (K, W, P, D), sequences from five different operators were recorded, and from three operators for the last two locations (G, M) [1]. Sample images from the dataset are shown in Fig. 6. Following the gaze tracker calibration, the operator moved freely between the locations performing verbally-communicated tasks (Tab. 1). Two sequences were recorded for each operator.

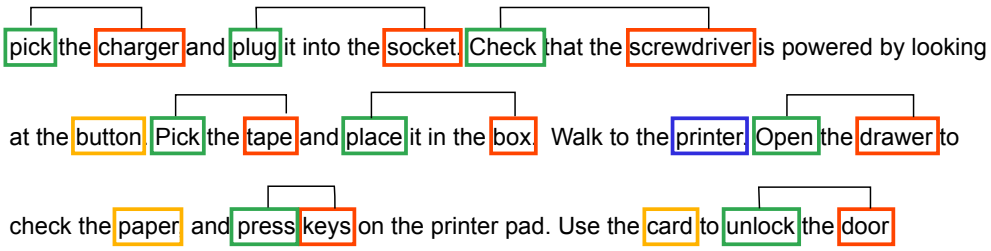


Figure 7: Example showing how the ground truth for TROs and MOIs was obtained from subject's narrations. Ground-truth TROs narrated by more than 50% of subjects are framed in red, compared to less-frequent subjects (orange). Location names are ignored (blue). The verb-noun combinations are used to ground-truth MOIs (green). The narrations are released with the dataset.

The operators were then asked to watch the videos, and write down a narration of what they have performed. Narrations were stemmed manually to match nouns

[1] Dataset available at: `http://www.cs.bris.ac.uk/~damen/BEOID`

	Number of sequences	Sequence length		Tracked (%)		Gaze Fixations (%)	
		μ	σ	μ	σ	μ	σ
K	10	1905	386	69.4	9.1	58.9	11.1
	Prepare coffee using the machine, place the cup on the mat and add sugar [tap, coffee machine, heat mat, cutlery drainer], (cup, sugar jar)						
W	10	1221	194	78.3	12.4	61.9	18.1
	Plug the screwdriver for charging and place the tape in the red box [Socket, Box], (screwdriver, charger, tape)						
P	10	596	77	75.8	13.3	70.5	14.1
	Check the printer is loaded with paper manually and using the keypad [drawer, keypad]						
D	10	303	83	71.8	15.8	56.2	14.7
	Go through the locked door [door lock, door handle]						
G	6	5183	482	76.4	9.0	66.7	11.0
	Use the treadmill and the bicycle next to it [treadmill panel, bicycle panel]						
M	6	2059	624	24.5	16.2	14.6	15.2
	Adjust the seat, chest pad and weight then use the machine [seat adjuster, pad adjuster, weight adjuster]						

Table 1: For the six locations, the number of sequences, average number of frames, percentage of tracked frames, percentage of gaze fixations, as well as the verbally communicated tasks, fixed "[]" and movable "()" ground-truth TROs.

and verbs which are semantically identical (e.g. adaptor vs. charger, pick vs. retrieve). Nouns narrated by more than 50% of the operators represent the twenty ground-truth TROs. Narrated verb-noun combinations are labelled as MOIs. Objects varied between having a single MOI (e.g. door handle: open) and up to three different usage methods (e.g. sugar jar: pick, put, get sugar). Figure 7 shows an example of how the narrations were used to generate the ground-truth TROs and MOIs.

For each location, a map of the environment is built using Parallel Tracking and Mapping (PTAM) [27]. A 3D bounding box around each object is manually labelled for evaluation. For moveable objects, their different locations are ground-truthed.

Parameters: In all results, the image parts $\Omega(I_t)$ were fixed to a window size of 200×200 pixels, This corresponds to 19.3° visual angles in the scene camera. To calculate appearance descriptors, $\Omega(I_t)$ is divided into 10×10 non-overlapping patches for calculating HOG descriptors. In offline processing, the number of words in BoW representation is set to 200. In calculating the BD index, $K = [2..2No_{gt}]$ (Eq. 4) where No_{gt} is the number of ground-truth objects. In online TRO discovery, ξ was set to 40 frames which corresponds to 1333ms of attention.

Results for discovering TROs: The results of offline and online TRO discovery are compared to the established ground-truth. The clusters' 3D bounding boxes are compared to ground-truth bounding boxes and the PASCAL overlap criteria

		w	clustering		Without Attention Ω_c			With Attention Ω_g		
					app	pos	both	app	pos	both
Offline	Known K	1	k-means	Recall	50.5	55.0	60.0	55.0	80.0	80.0
				Precision	52.6	61.1	66.7	61.1	84.2	84.2
				F-1 Score	51.5	57.9	63.2	57.9	**82.0**	**82.0**
			Spectral	Recall	45.0	60.0	50.0	60.0	80.0	90.0
				Precision	47.4	66.7	58.8	60.0	80.8	90.0
				F-1 Score	46.2	63.2	54.0	60.0	80.4	**90.0**
		25	k-means	Recall	50.0	60.0	55.0	60.0	85.0	85.0
				Precision	52.6	70.6	64.7	60.0	89.5	89.5
				F-1 Score	51.3	64.9	59.5	60.0	**87.2**	**87.2**
			Spectral	Recall	50.0	60.0	55.0	70.0	90.0	90.0
				Precision	55.6	66.7	57.9	73.7	90.0	94.7
				F-1 Score	52.7	63.2	56.4	71.8	90.0	**92.3**
	DB Index	1	k-means	Recall	35.0	40.0	40.0	55.0	65.0	65.0
				Precision	50.0	40.0	44.4	40.7	59.1	61.9
				F-1 Score	41.2	40.0	42.1	46.8	61.9	**63.4**
			Spectral	Recall	50.0	65.0	60.0	65.0	85.0	90.0
				Precision	41.7	54.2	52.2	41.9	68.0	75.0
				F-1 Score	45.5	59.1	55.8	51.0	75.6	**81.8**
		25	k-means	Recall	60.0	40.0	45.0	60.0	65.0	70.0
				Precision	44.4	42.1	52.9	42.9	59.1	63.6
				F-1 Score	51.0	41.0	48.6	50.0	61.9	**66.7**
			Spectral	Recall	70.0	75.0	60.0	70.0	80.0	95.0
				Precision	45.2	51.7	50.0	48.3	59.3	73.0
				F-1 Score	54.9	61.2	54.5	57.2	68.1	**82.6**
Online				Recall	26.7	7.2	40.0	73.3	13.3	85.0
				Precision	50.0	6.7	52.9	55.0	7.2	77.3
				F-1 Score	34.8	6.9	45.6	62.8	9.3	**81.0**

Table 2: Recall, precision and F1-score results for discovering TROs using different features, clustering methods, with/without attention and sliding window for the proposed offline and online TRO discovery methods.

(in 3D) of 20% indicates a true-positive. This is because the viewed positions don't typically cover the full extent of the object. Table 2 shows the complete set of results. In offline TRO discovery, two clustering methods are compared - spectral clustering and k-means. Appearance and position features are used individually or combined, either for a single frame ($w = 1$) or a sliding window ($w = 25$). The image part mechanisms Ω_c and Ω_g are compared, where the latter crops an image around gaze fixations thus referred to as cropping 'with attention'. Estimating the number of clusters using the Davies-Bouldin (DB) index is compared to knowing the number of clusters *a priori* (ref. *Known K*). For online results, the best precision for the highest recall is reported as the parameters

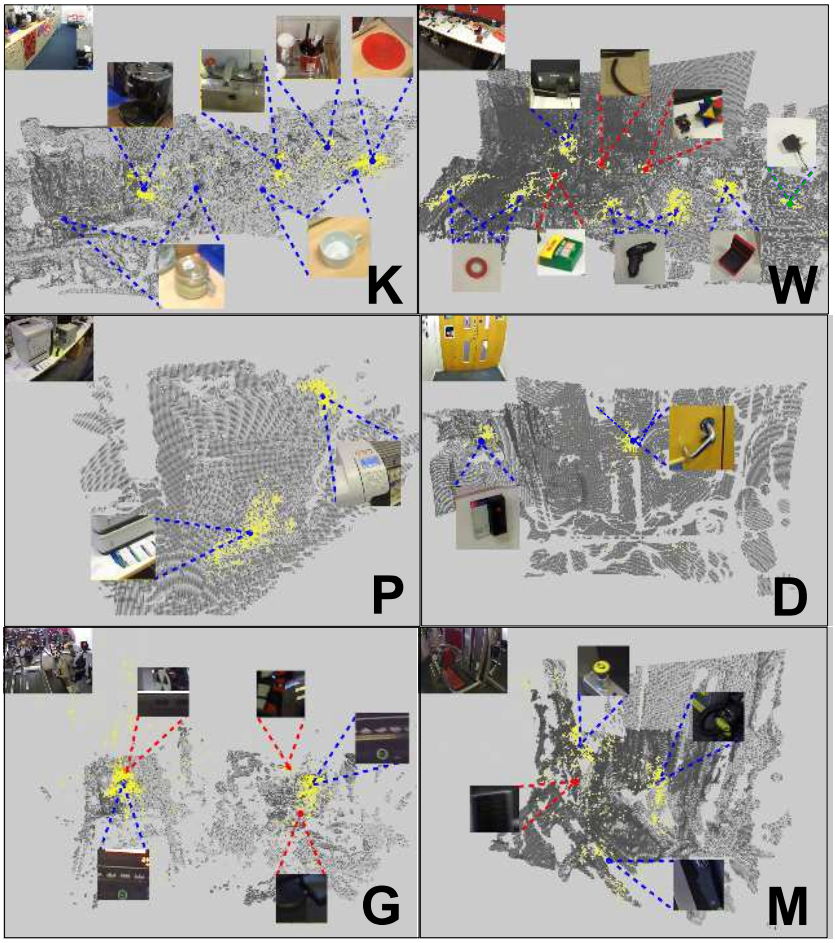


Figure 8: Discovered TROs (offline - appearance, position, attention, spectral clustering, $w = 25$ and DB index (i.e. number of objects is unknown)). An overview of the locations is shown at the top. Blue dots represent true-positive (19 objs), red dots represent false positive (7 objs) and green dots represent false negative (1 obj).

$(\epsilon_1, \epsilon_2, \epsilon_3)$ are varied.

Table 2 shows that the best offline results are obtained using spectral clustering, combining appearance and position, with attention and over a sliding window. Using Davies-Bouldin (DB) index, 95% of the TROs were retrieved with 73% precision. These discovered TROs are shown in Fig. 8. If the number of clusters was known *a priori*, 90% of TROs would be discovered with 94% precision. This is because the optimal number of clusters using DB index was higher than ground-truth K, resulting in one more correct object and several false positive clusters. In

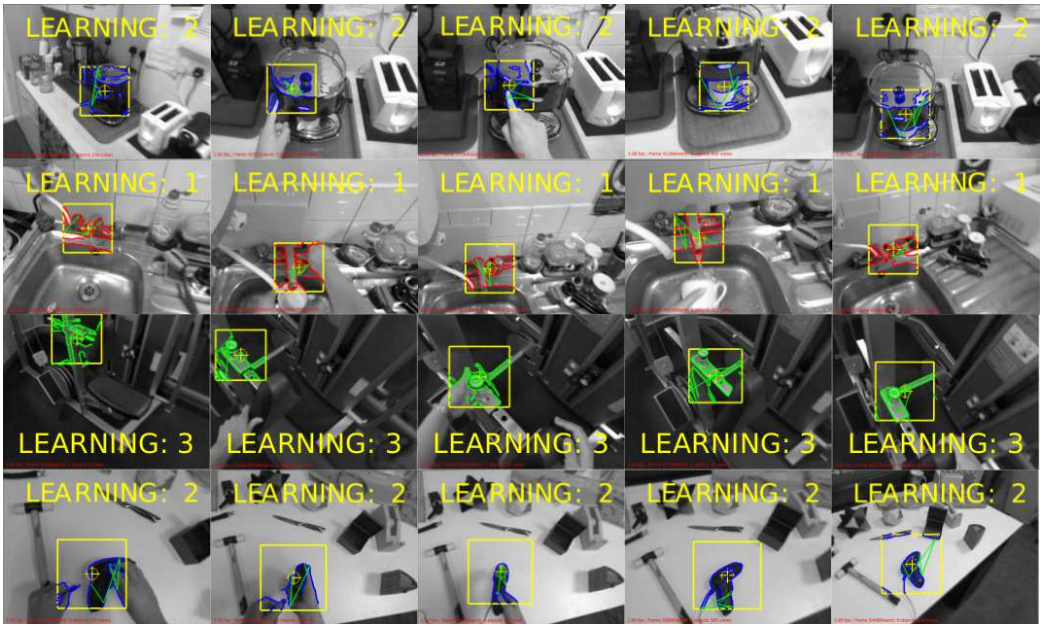
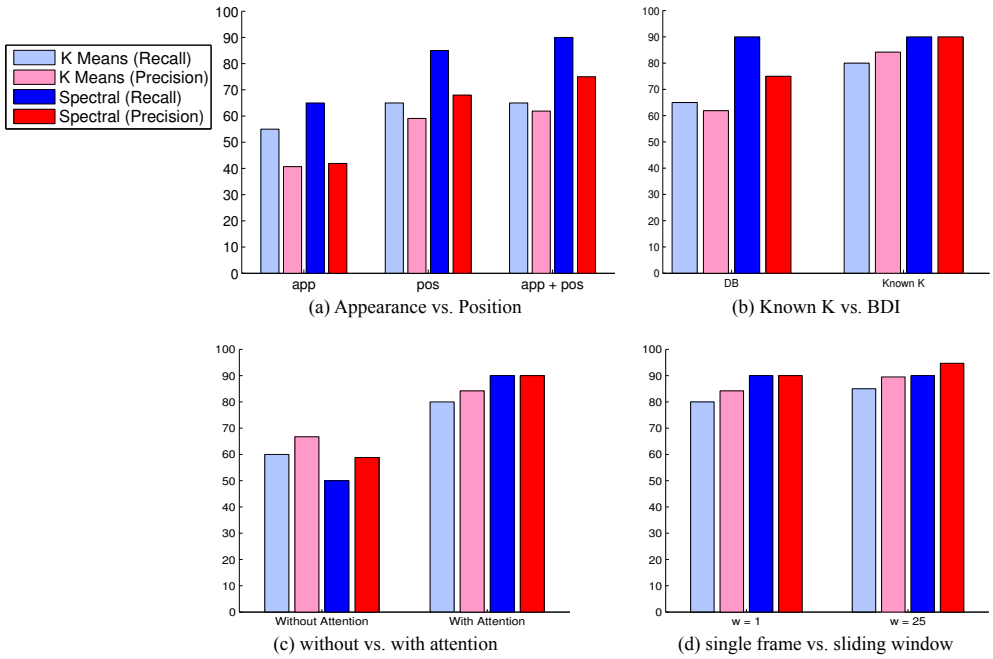


Figure 9: Learnt views from training sequences of multiple users for a variety of objects: coffee machine, tap, seat adjuster and screwdriver.

online TRO discovery, attention significantly improves the results as the chance of ξ consecutive similar image parts increases. Interestingly, when combining appearance and position, 85% of the objects were retrieved with 77% precision (F-1 score = 81%) showing the potentials of the scalable algorithm. Examples of learnt views for the discovered objects can be found in Fig 9.

Fig. 10 highlights several conclusions from the results of offline TRO discovery: (a) shows that for [DB, attention, $w = 1$] position achieves better than appearance when used solely. This is because most of the objects in our dataset (15/20) are fixed objects. As expected, adding appearance information increases the precision as this clusters instances of moveable objects into a single cluster. Fig. 10 (b) shows that DB index achieves the same recall as Known K when using spectral clustering [app+pos, attention, $w = 1$]. Precision increases when K is known - i.e. smaller discarded clusters actually do not represent TROs. Fig. 10 (c) shows the importance of within-image attention [app+pos, KnownK, $w = 1$]. A significant drop in recall is observed when the information is gathered around the image centre rather than gaze fixations. Fig. 10 (d) shows that a sliding window gives a slight improvement in performance.

Results for discovering MOIs For each discovered object, the usage snippets



(a) Appearance vs. Position (b) Known K vs. BDI

(c) without vs. with attention (d) single frame vs. sliding window

Figure 10: **(a)** appearance (app) vs position (pos) and their combination (app+pos) using spectral vs. k-means clustering using DB index. **(b)** Using app+pos, DB index vs. known number of clusters. **(c)** For app+pos+knownK, parts centred around centre of image vs. gaze fixations. **(d)** Single-frame vs. sliding window representations.

longer than $\xi = 1s$ are used to build a usage model. On average, 16.6 usage snippets are extracted for each TRO ($\sigma = 7.4$). Notice that these snippets are extracted automatically based on the discovered TRO. The example shown here is from the online discovery of TROs for the object *tap*. We vary the threshold λ to accept $p(MOI_j)$ (Eq. 11) and plot recall-precision curves. A cluster is true-positive if its representative snippet matches one ground-truth MOI; a duplicate match for the same ground-truth MOI is a false-positive.

We compare using position, appearance and motion features with a temporal pyramid as well as their combination (Fig. 11). The figure shows that while position information benefits from the temporal pyramid, achieving its highest performance at $L = 3$, motion information achieves its best information without using a temporal pyramid $L = 1$. We believe this is due to the various speeds at which people perform the motion. As anticipated, motion information solely is capable of distinguishing the various modes of interaction with the same object. Using the combination of features and $\lambda = 0.2$ (Eq. 11), the approach is able to

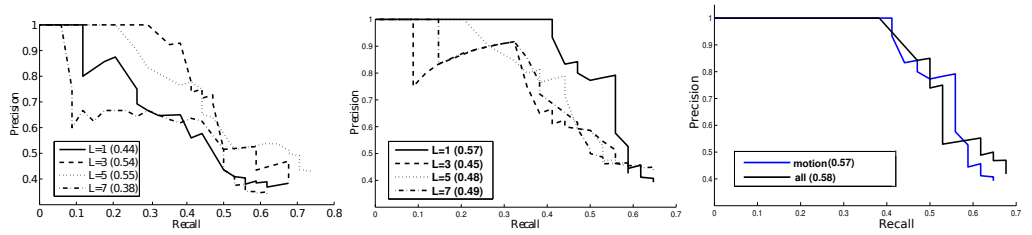


Figure 11: For position (left), temporal pyramid (L=5) performed best, while motion (right) performed best on L=1. When using motion only versus combining all features at their best temporal pyramid level, a minor improvement is observed.

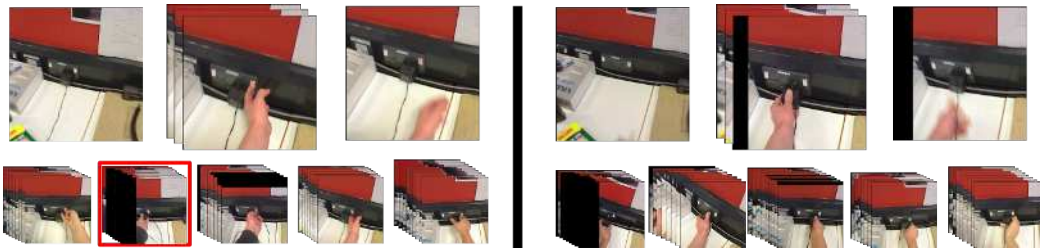


Figure 12: For the 'socket', the two common MOIs ('switching', 'plugging') are found (left & right). The representative *usage snippet* is shown (up) with the other snippets in the same cluster (below) - only one snippet is incorrectly clustered (shown in red).

discover meaningful MOIs. Figure 12 shows an example of the method successfully discovering two MOIs for the 'socket'. Given 10 automatically extracted usage snippets, snippets representing the 'switch' and 'plug' MOIs are separated, with a single snippet incorrectly clustered. Notice that the motion descriptors are used for clustering without any discriminative tuning to achieve this separation. Similarly, Fig. 13 shows further discovered MOIs for the sugar jar and the door handle. The proposed method is able to discover objects with a single as well as multiple MOIs. For the sugar jar, the representative usage snippets show the MOIs 'get sugar', 'put' and 'pick' separated. For the door handle a single MOI is considered common with smaller clusters discarded as spurious.

Results for Graph of Object Interactions: With the discovered TROs, we trained the graphs representing the interactions. The initial link α was set as 0.05. The generated graphs for the Kitchen (K) and weight Machine (M) sequences are presented in Figure 14. Notice the strong causal links: *coffee machine→heat mat*, *tap→coffee machine*, *sugar jar→heat mat*, *seat adjuster→pad adjuster* all being meaningful strong links between interactions with these objects in the dataset.

Results for Assistive Mode: While we do not test the assistive mode with users

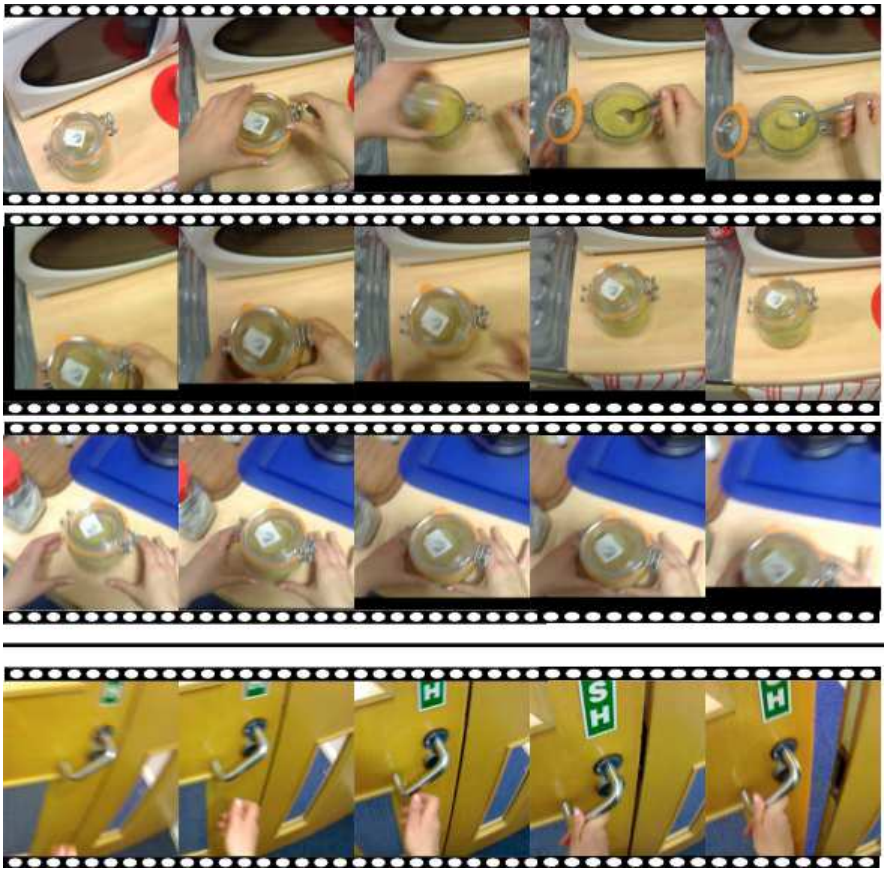


Figure 13: For TRO ‘jar’, 3 MOIs are discovered (‘get sugar’, ’put’, ’pick’). For the handle, one MOI is discovered. Frames from the representative snippets are shown.

to evaluate the ‘usefulness’ of the provided usage snippets or the recommendation for the object to use next, we qualitatively assess the ability of the assistive mode to provide meaningful help snippets. In the assistive mode, we use a leave-one-out; for every operator, TROs are discovered and common MOIs are found from sequences of other operators. In the assistive mode, when a discovered TRO is detected, an insert is shown indicating a suggestive way of how the object can be used and what object to use next. We show results from the two recognition methods, first employing the position models to predict the object being used, then employing the appearance models.

Fig 15 shows a sequence of object interactions in the assistive mode, using location models to recognise TROs. When the user fixates at a discovered TRO,

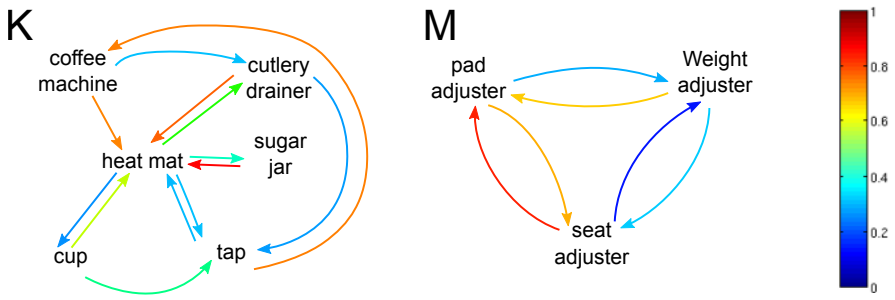


Figure 14: Graph of TRO interactions for two locations kitchen (K) and gym machine (M). Weighting scores of edges in the graph are portrayed as a heat colour map, and edges of weight $\leq \alpha$ are not shown.

a usage snippet indicating how to use that object is recommended along with the object to be used next. The figure also shows links (coloured using the heat map in Fig. 14) to indicate weight of the edges in the object interactions graph. When the TRO *heat mat* is recognised (Fig 15a), a usage snippet is shown recommending a cup to be placed on the mat. The next object to be used is thought to be the TRO *drainer*. The operator indeed moves towards the drainer (Fig 15b), and the drainer is recognised (Fig 15c). The recommended usage is to pick a cutlery and the suggested next object is the heat mat. The harvested view of the heat mat is that with the cutlery being used. Though this is automatically chosen, it is extracted from the set of usage snippets that follow using the drainer. The attention is indeed shifted to the heat mat (Fig 15d).

Next, we use the real-time texture-minimal scalable detector from [7] to recognise TROs, due to its light-weight computational load that makes it amendable to wearable systems [4, 19]. Note that when using the appearance models, a map of the environment would not be needed in the assistive mode. A *help snippet* is displayed each time a new object is recognised. We showcase video help guides using inserts on a pre-recorded video. These could in principle be shown on a head-mounted display, but is not considered in this study. Figure 16 shows frames from the help videos and a full sequence is available [2]. Recall that these inserts are *extracted, selected and displayed* fully automatically. This assistive mode presents a possible application for unsupervised discovery of TROs and their MOIs. We believe other potential applications could be explored.

[2] `https://youtu.be/vUeRJmwm7DA`

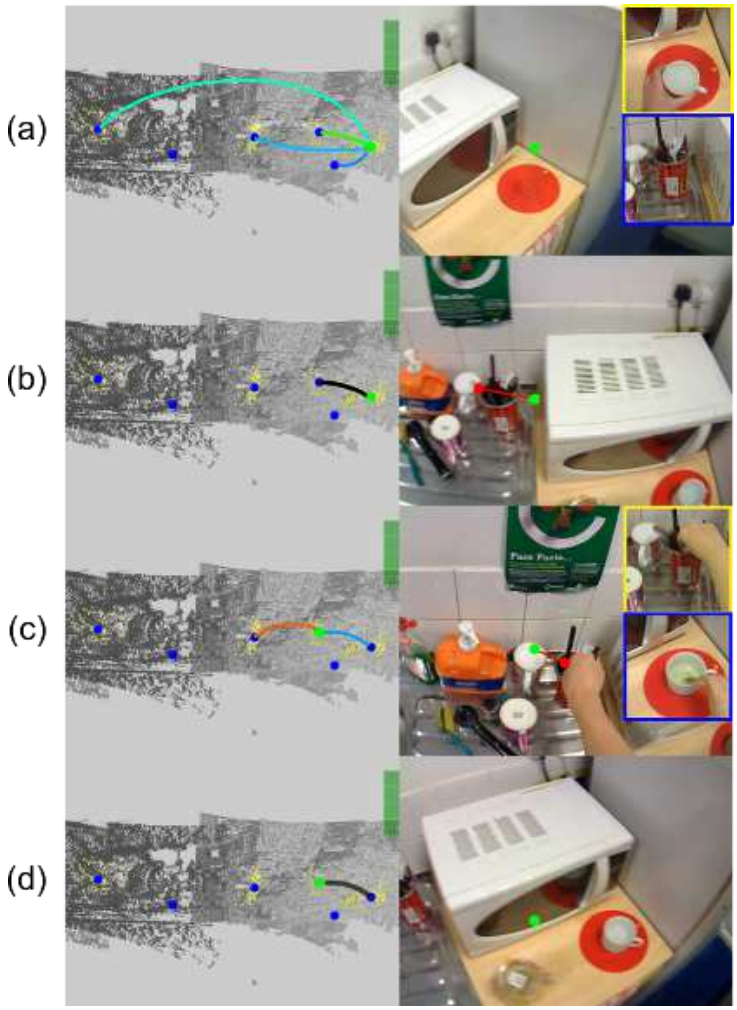


Figure 15: For a pre-built map of the environment (left) and egocentric images with tracked gaze (right), the position (green-dot) is used to recognise TRO, a usage snippet is inserted (yellow-framed insert) along with the object to be used next (blue-framed insert). The recommendations are inserted everytime a different TRO is recognised.

6. Building 3D Models of Task-Relevant Objects

To build a three dimensional representation of the object O_k, we adapt the work of [34] so it does not require the detection of keyframes and it uses input from multiple users. Given a sparse map of the environment, the 3D points-of-regard are found by back-projecting the rays connecting the camera to the image

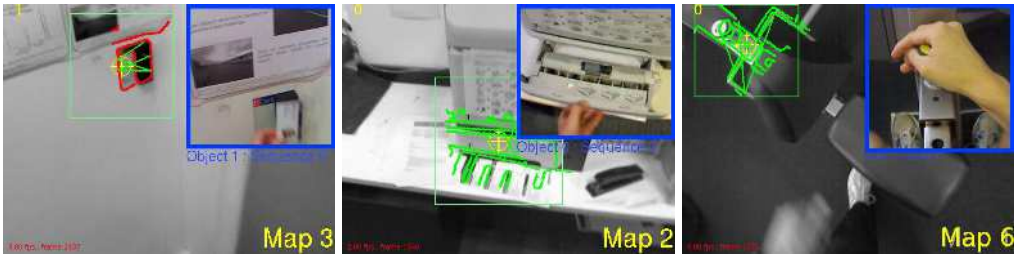


Figure 16: In the assistive mode, when a TRO is recognised using, usage snippet is inserted (blue-frame) showing the most relevant common MOI based on the initial appearance.

part. These are used as seeds for super-pixel segmentation. The method uses outlier removal to reduce the error in volume estimation. In this work, we exploit 3D position information to generate textured three-dimensional models of the TROs. Despite not being perfect models, due to the fact that they are created during task performance, the resulting models are useful visualisations of what objects the system has discovered. Ultimately, having a 3D model could facilitate applications such as augmented reality guidance.

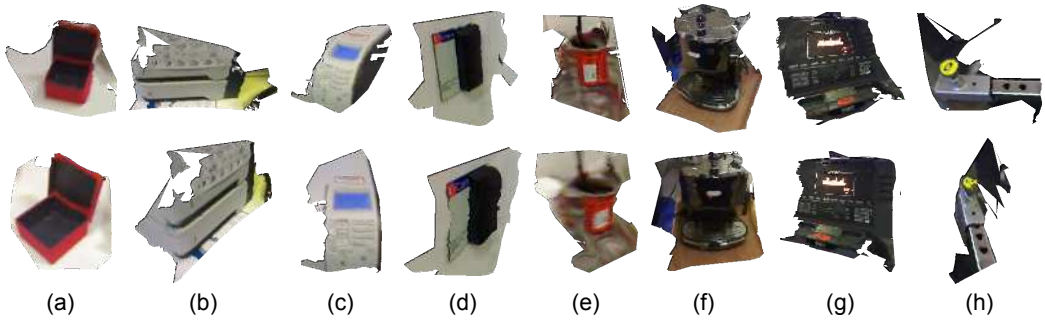


Figure 17: Textured three-dimensional models (two views each) for eight discovered TROs.

The accuracy of the model relies on whether it has been viewed from multiple views by the users. The importance of sequences from multiple users is particularly noticed when attempting to build these approximate 3D models. Figure 17 presents three-dimensional models for eight discovered objects. Note that the method is capable of discovering and representing small-sized (a,d,e,h) as well as larger objects (b,f,g).

7. Conclusion and Future Work

In this work, we present an approach for discovering task relevant objects and their common modes of interaction from multi-user egocentric video, *fully automatically*. We compare appearance, position and motion features, along with gaze fixations to indicate attention, for the discovery. For an unknown number of objects, the approach relies on clustering along with a clustering evaluation measure. We compare offline clustering to an online algorithm that iteratively refines and updates the clusters. Both approaches are able to discover fixed objects, such as the sink or a socket as well as moveable objects such as a cup or a sugar jar, as the approaches combine location and appearance features.

On a newly introduced and published egocentric dataset that spans six locations, detailed results show that the offline approach achieved highest performance (F-1 score of 92.3%) using spectral clustering over a sliding window, by combining appearance and position features with a gaze fixation attention model. The online approach achieves an F-1 score of 81%. Moreover, for each discovered object, various usage snippets have been automatically extracted and clustered using motion features to discover modes of interaction. First-order Markovian assumption is also employed to build a graph of possible interactions, based on sequences of interactions performed by multiple operators.

Discovered task-relevant objects can be used for providing assistance to users. As opposed to approaches that require manual authoring of assistance for an object or a task, the assistance proposed here is unsupervised. Triggered by gazing at the object to be used, the appearance model would recognise a previously discovered task relevant object. Video-based snippet guidance can then suggest a mode of interaction, given the current status of the object. This is particularly important for objects that change appearance resulting in varying functionality. Moreover, a graph of object interactions can be employed to suggest an object to be used next. This assistance is useful for objects that are often used in consecutive order (e.g. the sink and the drainer or the door lock and the door handle).

The paper provides detailed comparative evaluation for baseline appearance and motion features. State-of-the-art motion (e.g. dense trajectories) and appearance (e.g. convolutional neural networks for tuned discriminative features) features could be investigated. The approach fails to discover objects with very short gaze fixation durations. This is particularly true for objects the user picks up as soon as they are observed. The scalability of the method to discover modes of interaction from multiple tasks requires further research. Currently objects with up to 3 common modes of interaction have been tested.

The paper highlights the importance of attention for discovering task-relevant objects. In this work, we use gaze fixations as a mechanism for detecting attention, both temporally and spatially. Currently, only a few wearable setups offer wearable gaze tracking. Approaches that estimate attention could alternatively be deployed. Our recent work [33] has detailed a method to estimate attention, both temporally and spatially, from Inertial Unit Measurements (IMU). Alternatively, a method to estimate a visual attention map from the visual flow in an image has been proposed by Matsuo et al [38]. Testing the ability of estimated attention to discover task relevant objects and their modes of interaction is one future direction.

While this paper provides a complete framework that bridges the gap between unsupervised object discovery and video-based guidance with promising preliminary results, it aims to initiate further research and discussions, particularly related to the usefulness of automatically extracted video guides for human operators and/or autonomous systems, the importance of attention information in egocentric video analysis and more advanced techniques towards discovering modes of interactions for everyday objects.

Acknowledgement We would like to thank Pished Bunnun, Osian Haines and Andrew Calway for their input on previous iterations of parts of this work.

References

[1] Betancourt, A., Morerio, P., Regazzoni, C., Rauterberg, M., 2015. The evolution of first person vision methods: A survey. IEEE Trans. on Circuits and Systems for Video Technology 25 (5).

[2] Bleser, G., Almeida, L., Behera, A., Calway, A., Cohn, A., Damen, D., Dominatedes, H., Gee, A., Gorecky, D., Hogg, D., Kraly, M., Macaes, G., Marin, F., Mayol-Cuevas, W., Miezal, M., Mura, K., Petersen, N., Vignais, N., Santos, L., Spaas, G., Stricker, D., 2013. Cognitive workflow capturing and rendering with on- body sensor networks (cognito). German Research Center for Artificial Intelligence, DFKI Research Reports (RR).

[3] Bolanos, M., 2015. Ego-object discovery in lifelogging datasets. Master's thesis, Universitat de Barcelona.

[4] Bunnun, P., Damen, D., Calway, A., Mayol-Cuevas, W., 2012. Integrating 3D object detection, modelling and tracking on a mobile phone. In: Int. Symp. on Mixed and Augmented Reality (ISMAR).

[5] Collet, A., Xiong, B., Gurau, C., Hebert, M., Srinivasa, S., 2013. Exploiting domain knowledge for object discovery. In: Int. Conf. on Robotics and Automation (ICRA).

[6] Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition (CVPR).

[7] Damen, D., Bunnun, P., Calway, A., Mayol-Cuevas, W., 2012. Real-time learning and detection of 3D texture-less objects: A scalable approach. In: British Machine Vision Conference (BMVC).

[8] Damen, D., Haines, O., Leelasawassuk, T., Calway, A., Mayol-Cuevas, W., 2014. Multi-user egocentric online system for unsupervised assistance on object usage. In: ECCV Workshop on Assistive Computer Vision and Robotics (ACVR).

[9] Damen, D., Leelasawassuk, T., Haines, O., Calway, A., Mayol-Cuevas, W., 2014. You-do, I-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In: British Machine Vision Conference (BMVC).

[10] Davies, D., Bouldin, D., 1979. A cluster separation measure. Pattern Analysis and Machine Intelligence (PAMI) 1 (2), 224–227.

[11] De Beugher, S., Ichiche, Y., Brone, G., Geodeme, T., 2012. Automatic analysis of eye-tracking data using object detection algorithms. In: Workshop on Perasive Eye Traking and Mobile Eye-based Interaction (PETMEI).

[12] Fathi, A., Li, Y., Rehg, J., 2012. Learning to recognize daily actions using gaze. In: European Conference on Computer Vision (ECCV).

[13] Fathi, A., Rehg, J., 2013. Modeling actions through state changes. In: Computer Vision and Pattern Recognition (CVPR).

[14] Fathi, A., Ren, X., Rehg, J., 2011. Learning to recognise objects in egocentric activities. In: Computer Vision and Pattern Recognition (CVPR).

[15] Gonzalez-Diaz, I., Boujut, H., Buso, V., Benois-Pineau, J., Domenger, J., 2014. Saliency-based object recognition in video. HAL Archives.

[16] Goto, M., Uematsu, Y., Saito, H., Senda, S., Iketani, A., 2010. Task support system by displaying instructional video onto ar workspace. In: Int. Symposium on Mixed and Augmented Reality (ISMAR).

[17] Hashimoto, Y., Kondo, D., Yonemur, T., Iizuka, H., Ando, H., Maeda, T., 2011. Improvement of wearable view sharing system for skill training. In: Int. Conference on Artificial Reality and Telexistence.

[18] Herbst, E., Henry, P., X, R., Fox, D., 2011. Toward object discovery and modeling via 3-D scene comparison. In: Int. Conf. on Robotics and Automation (ICRA).

[19] Hodan, T., Damen, D., Mayol-Cuevas, W., Matas, J., 2015. Efficient textureless object detection for augmented reality guidance. In: Mixed and Augmented Reality Workshops (ISMARW).

[20] Johansson, R., Westling, G., Backstrom, A., Flanagan, J., 2001. Eye-hand coordination in object manipulation. The Journal of Neuroscience 21 (17).

[21] Kang, H., Hebert, M., Efros, A., Kanade, T., 2015. Data-driven objectness. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 37 (1).

[22] Kang, H., Hebert, M., Kanade, T., 2011. Discovering object instances from scenes of daily living. In: Int. Conference on Computer Vision (ICCV).

[23] Kang, S., Ikeuchi, K., 1994. Determination of motion breakpoints in a task sequence from human hand motion. In: Int. Conference on Robotics and Automation (ICRA).

[24] Karpathy, A., Miller, S., Fei-Fei, L., 2013. Object discovery in 3d scenes via shape analysis. In: Int. Conf. on Robotics and Automation (ICRA).

[25] Kim, G., Faloutsos, C., Herbert, M., 2008. Unsupervised modeling of object categories using link analysis techniques. In: Computer Vision and Pattern Recognition (CVPR).

[26] Kitani, K., Okabe, T., Sato, Y., Sugimoto, A., 2011. Fast unsupervised ego-action learning for first-person sports videos. In: Computer Vision and Pattern Recognition (CVPR).

[27] Klein, G., Murray, D., 2007. Parallel Tracking and Mapping for Small AR Workspaces. In: Int. Sym. on Mixed and Augmented Reality (ISMAR).

[28] Laboratories, A. S., 2011. Mobile Eye-XG.
URL http://www.asleyetracking.com/

[29] Land, M., 2006. Eye movements and the control of actions in everyday life. Progress in Retinal and Eye Research 25 (3).

[30] Land, M., Mennie, N., Rusted, J., 1999. The roles of vision and eye movements in the control of activities of daily living. Perception 28 (11).

[31] Laptev, I., 2005. On space-time interest points. Int. Journal of Computer Vision (IJCV) 64 (2).

[32] Lee, Y., Ghosh, J., Grauman, K., 2012. Discovering important people and objects for egocentric video summarization. In: Computer Vision and Pattern Recognition (CVPR).

[33] Leelasawassuk, T., Damen, D., Mayol-Cuevas, W., 2015. Estimating visual attention from a head mounted imu. In: International Symposium on Wearable Computers (ISWC).

[34] Leelasawassuk, T., Mayol-Cuevas, W., 2013. 3D from looking: Using wearable gaze tracking for hands-free and feedback-free object modelling. In: Int. Sym. on Wearable Computers (ISWC).

[35] Li, Y., Fathi, A., Rehg, J., 2013. Learning to predict gaze in egocentric video. In: Int. Conf. on Computer Vision (ICCV).

[36] Lu, Z., Grauman, K., 2013. Story-driven summarization for egocentric video. In: Computer Vision and Pattern Recognition (CVPR).

[37] Mason, J., MArthi, B., Parr, R., 2012. Object disappearance for object discovery. In: Int. Conf. or Intelligent Robots and Systems (IROS).

[38] Matsuo, K., Yamada, K., Ueno, S., Naito, S., 2014. An attention-based activity recognition for egocentric video. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[39] Mayol, W., Tordoff, B., Murray, D., 2009. On the choice and placement of wearable vision sensors. IEEE Trans. on Systems Man And Cybernetics 39.

[40] Mayol-Cuevas, W., Davison, A., Tordoff, B., Murray, D., 2004. Applying active vision and SLAM to wearables. In: International Symposium on Robotics Research.

[41] Mayol-Cuevas, W., Murray, D., 2005. Wearable hand activity recognition for event summarization. In: IEEE Int Symposium on Wearable Computers (ISWC).

[42] Ng, A., Jordan, M., Weiss, Y., 2002. On spectral clustering: analysis and an algorithm. In: Advances in Neural Information Processing Systems (NIPS).

[43] Nguyen, T., Nebel, J., Florez-Revuelta, F., 2016. Recognition of activities of daily living with egocentric vision: A review. Sensors 16 (1).

[44] Ogaki, K., Kitani, K., Sugano, Y., Sato, Y., 2012. Coupling eye-motion and ego-motion features for first-person activity recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW).

[45] Petersen, N., Stricker, D., 2012. Learning task structure from video examples for workflow tracking and authoring. In: International Symposium on Mixed and Augmented Reality (ISMAR).

[46] Pirsiavash, H., Ramanan, D., 2012. Detecting acitivites of daily living in first-person camera views. In: Computer Vision and Pattern Recognition (CVPR).

[47] Ren, X., Gu, C., 2010. Figure-ground segmentation improves handled object recognition in egocentric video. In: Computer Vision and Pattern Recognition (CVPR).

[48] Ren, X., Philipose, M., 2009. Egocentric recognition of handled objects: Benchmark and analysis. In: Computer Vision and Pattern Recognition Workshop (CVPRW).

[49] Russell, B., Efros, A., Sivic, J., Freeman, W., Zisserman, A., 2006. Using multiple segmentations to discover objects and their extent in image collections. In: Computer Vision and Pattern Recognition (CVPR).

[50] Salvucci, D., Goldberg, J., 2000. Identifying fixations and saccades in eye-tracking protocols. In: Sym. on Eye Tracking Research & Applications.

[51] Sanders, B., Nelson, R., Sukthankar, R., 2002. A theory of the quasi-static world. In: Int. Conf. on Pattern Recongition (ICPR).

[52] Somanath, G., Mv, R., Metaxas, D., Kambhamett, C., 2009. D - clutter: Building object model library from unsupervised segmentation of cluttered scenes. In: Computer Vision and Pattern Recognition (CVPR).

[53] Sun, L., Klank, U., Beetz, M., 2009. EyeWatchMe - 3D hand and object tracking for inside out activity analysis. In: Computer Vision and Pattern Recognition Workshop (CVPRW).

[54] Sundaram, S., Mayol-Cuevas, W., 2009. High level activity recognition using low resolution wearable vision. In: First Workshop on Egocentric Vision, Computer Vision and Pattern Recognition (CVPRW).

[55] Sundaram, S., Mayol-Cuevas, W., 2012. What are we doing here? egocentric activity recognition on the move for contextual mapping. In: Int. Conf. on Robotics and Automation (ICRA).

[56] Takemura, K., Kohashi, Y., Suenaga, T., Takamatsu, J., Ogasawara, T., 2010. Estimating 3D point-of-regard and visualizing gaze trajectories under natural head movements. In: Sym. on Eye-Tracking Research & Applications (ETRA).

[57] Tuytelaars, T., Lampert, C., Blaschko, M., Buntine, W., 2010. Unsupervised object discovery: A comparison. Int. Journal on Computer Vision (IJCV).

[58] Yan, Y., Ricci, E., Liu, G., Sebe, N., 2014. Recognizing daily activities from first-person videos with multi-task clustering. In: Asian Conference on Computer Vision (ACCV).